Data management plan covering common long-term data for the MILC and Fermilab Lattice collaborations

Fermilab Lattice and MILC Collaborations

March 6, 2021

Abstract

This data management plan (DMP) covers long term data common to MILC and Fermilab Lattice collaboration projects. Among the data products covered are the MILC HISQ gauge configuration ensembles, Dirac operator eigenvector sets, and hadronic correlator data produced in analysis campaigns. The gauge ensembles and hardronic correlator data are scientifically valuable long term since they form the basis of published analyses. The Dirac operator eigenvector sets are cost effective to store since they are used repeatedly to accelerate solves for physical mass light quark propagators. Individual MILC and Fermilab projects will cover their specific data management plans separately.

1 Plan revision history

March 6, 2021 Plans updated to include configuration generation in 2020.

June 1, 2020 Initial revision.

March 5, 2021 Update.

2 Contact information

- *Provide the name and affiliation of the projects PI or main point of contact. Also provide the name and affiliation of the project's data managers. The data managers and the POC are the project members having primary responsibility for making decisions regarding the data covered in this plan.*

Role	Name	Affiliation
PI	Carleton DeTar	University of Utah
PI	Andreas Kronfeld	Fermilab
Data manager	Steve Gottlieb	Indiana University, Bloomington
Data manager	James Simone	Fermilab

3 Data types

- *DMPs for science projects applying for long-term storage should tabulate the amount of data expected to be produced in each of the next three years going forward. Include in the table the year the data will be produced, a concise unique identifier, data types (e.g., gauge field, eigenvectors, etc.), expected data set size produced in that year, number of files, and file sizes. For existing datasets, add a single line to the table listing the years it was produced, e.g., 2016–2019 and the other statistics. Please add any additional comments following the table.*

Data sets discussed here include the MILC HISQ gauge configuration ensembles, Dirac eigenvector sets, and hadronic correlator data produced in analysis campaigns.

The HISQ gauge configuration ensembles are of high scientific value as they are common to the many physics analyses conducted by the Fermilab Lattice and MILC collaborations. The ensembles are publicly available and are used by other collaborations world wide as well. The collection of ensembles includes many world-class features: a wide range of lattice spacings from 0.15 fm down to 0.03 fm, four flavors of sea quarks, ensembles with physical mass sea quarks, and a variety of ensembles with parameter choices useful to characterize systematic effects. Table 1 describes existing ensemble data sets as of March 6, 2021. These data sets have already been written to tape.

Table 1: HISQ gauge ensembles as of March 6, 2021. The most important gauge ensembles that are the basis of the Fermilab Lattice and MILC analysis campaigns.

created	a [fm]	data set	file count	File size [GB]	total [TB]
pre-2020	0.15	$m_s/5$	1021	0.057	0.06
pre-2020		$m_s/10$	1000	0.191	0.19
pre-2020		physical	1000	0.453	0.45
pre-2020		physical retune	10026	0.453	4.54
pre-2020		phys. $n_f = 1+1+1+1$	4949	0.453	2.24
pre-2020	0.12	$m_s/5$	1053	0.255	0.27
pre-2020		$m_s/10$	1000	1.18	1.18
pre-2020		physical	1020	2.038	2.08
2019–2020		physical retune	10000	2.038	20.38
pre-2020		$m_s/10$ small V	1020	0.255	0.26
pre-2020		$m_s/10$ large V	1030	1.18	1.22
pre-2020		unphysA	1020	0.255	0.26
pre-2020		unphysB	1020	0.255	0.26
pre-2020		unphysC	1020	0.604	0.62
pre-2020		unphysD	1020	0.604	0.62
pre-2020		unphysE	1020	0.604	0.62
pre-2020		unphysF	1020	0.604	0.62
pre-2020		unphysG	1020	0.604	0.62
pre-2020		unphysD $n_f = 1+1+1+1$	1186	0.604	0.72
pre-2020	0.088	$m_s/5$	1010	0.906	0.92
pre-2020		$m_s/10$	1001	3.058	3.06
2015–2020		physical	5442	7.248	39.44
pre-2020	0.057	$m_s/5$	1017	4.586	4.66
pre-2020		$m_s/10$	1249	10.872	13.58
2014–present		physical	2235	48.922	110.9
pre-2020	0.042	$m_s/5$	1301	14.496	18.86
2014–present		physical	519	247.669	128.54
pre-2020	0.03	$m_s/5$	1021	73.384	74.92
Totals:			56176		432.1

In 2020 we began generating new HISQ ensembles useful for understanding the sensitivity of our analyses to the sea quark masses. Table 2 is an estimate of our storage requirements for the new ensembles based on the target number of configurations. Only a fraction of the target numbers of configurations have been generated and written to tape at present.

Table 2: Projections for the number additional gauge configurations we expect to produce in the next few years. Counts of existing un-equilibrated and equilibrated configurations are given.

years	a [fm]	$(m'_l/m_s, m'_s/m_s)$	unequil.	equil.	target files	file size [GB]	total [TB]
2020 − 2021	0.088	$(1/5, 3/5)$	49	284	1050	0.906	0.95
2020 − 2021		$(1/10, 3/5)$			1050	3.058	3.21
2020 − 2021		$(1/10, 1/10)$	66	0	1075	3.058	3.29
2020 − 2021	0.057	$(1/5, 3/5)$	49	284	1050	4.586	4.82
2021 − 2022		$(1/10, 3/5)$			1050	10.872	11.42
2021 − 2022		$(1/10, 1/10)$			1075	10.872	11.69
Totals:					6350		35.38

The low lying eigenvectors of the Dirac operator are computationally expensive to compute. It is cost effective to store sets of low-mode eigenvectors for the light quark on the physical mass ensembles since costs are rapidly amortized by using the eigenvectors to accelerate repeated quark propagator solves. Table 3 lists the current sets on eigenvectors. We propose to continue generating eigenvector sets in 2021. Our storage projections are shown in Table 4.

Table 3: Status of low-lying Dirac eigenvector sets for the lightest quark as of 2021-02-27. A set of eigenvectors is computationally expensive to produce, hence, it is cost effective to store them for reuse. Costs are rapidly amortized by using the eigenvectors to accelerate repeated solves for the lightest mass quark propagators.

created	a [fm]	data set	file count	File size [GB]	total [TB]
2019	0.15	physical retune	1327	26.42	32.3
2019	0.12	physical retune	319	54.36	16.1
2019	0.088	physical	131	603.98	45.30
2020		physical	43	1207.96	79.4
Totals:			1820		183.5

Table 4: Low-lying Dirac eigenvector sets we expect to produce in the next USQCD allocation year.

year	a [fm]	data set	file count	File size [GB]	total [TB]
2021	0.15	physical retune	100	75.49	7.55
2021	0.088,	physical	300	603.98	181.19
Totals:			400		196.28

In Table 5 we list current storage requirements for the hadronic correlators produced by our analysis campaigns. These data are critical to preserve long term since they are the basis of the analyses appearing in our publications. Included with the milc outputs are the values of the hadron correlators and standard output from the application codes.

Table 5:

campaign	data format	file count	total size [TB]
all HISQ	milc outputs	43.6K	3.95
	sqlite	26	3.01
	hdf5	26	0.11
clover-hisq	milc outputs	60.5K	0.41
	sqlite	8	0.33
axial	milc outputs	1022.0K	0.51
g-2	milc outputs	47.2K	0.05
Totals:		1173.3K	8.39

4 Data and metadata standards

- *For the data types tabulated in the previous section, briefly describe the file format and the code bases that produced the data. Describe software interfaces that are able to read the files. Describe the nature, formats and location of the metadata describing the data.*

The MILC gauge ensembles (Tables 1 and 2) may be stored in the MILC v5 gauge configuration format or in the SciDAC ILDG format. Both formats are readable with the github:milc_qcd code. The SciDAC ILDG format is widely supported by many major LQCD code bases using the USQCD github:QIO library. The MILC code is able to translate between v5 and ILDG formats. The Dirac eigenvector files (Tables 3 and 4) are in SciDAC QIO format and the MILC code supports IO for such eigenvector files. Other application codes are able to access the eigenvector through the QIO API. The QIO library supports multiple levels of embedded metadata in XML format. Gauge and eigenvector file metadata encodes a description of the data: the QCD data type, the float precision, the lattice dimensions, and a creation timestamp. File names are decorated with additional metadata such as a unique ensemble identifier and the data's series and trajectory number within the ensemble.

The hadronic correlator data in Table 5 exists in a variety of formats. The primary representation is text output from the milc applications. Typically, the data are written to many small files at each incremental step of the campaign. The data are then packaged into compressed tarballs. The milc outputs are post processed and packaged in an sqlite embedded relational database. The sqlite data are not averaged or folded hence the sqlite contents is as general as the milc outputs. In addition the sqlite representation preserves relations among the data and metadata describing the correlators. The Library of Congress considers sqlite an archival data format. The sqlite data is further processed and written in hdf5 format which is the most convenient representation for input into the physics analysis. The latter step does do reductions averaging each configuration over multiple time sources and doing any reductions needed for the truncated solve variance reduction technique. We intend to backup the primary milc outputs data and secondary representations of the hadronic data to long-term tape storage at Fermilab. The primary milc outputs can then be removed from disk. We intend to maintain the secondary representations on disk allocated to Fermilab scientists and their collaborators.

5 Policies for access, sharing, and protection

- *Gauge ensembles generated using USQCD resources are made available to the USQCD community. Describe your project's plan for providing access and sharing within USQCD. Describe your collaboration's policies governing allowable usage of these data by other USQCD projects. Also describe any mechanisms in place to protect these data from unauthorized access.*

The HISQ gauge configurations are freely shared with other USQCD science projects. Projects having goals similar or in competition with MILC and Fermilab projects are asked to negotiate use and respect MILC's priority to publish first. JNS: *What is the official policy??* The HISQ gauge configuration ensembles and Dirac eigenvector sets are stored on tape at Fermilab. Subsets of the gauge configurations are cached on the Lustre filesystem attached to the Fermilab institutional cluster, wc.fnal.gov, under directory /lustre1/publicdata/milc/hisq. The Lustre directories are publicly readable by any USQCD member with an account on the cluster. Members of USQCD collaborations are able to contact a data manager of this plan to request that additional gauge configurations be cached from magnetic tape.

To guard against accidental data loss the ensembles are replicated on tape resources at the TACC facility.

6 Policies for re-use and redistribution

- *Describe your collaboration's limitations, policies and plans on sharing these data outside of USQCD.*

The HISQ gauge configurations are freely shared with other researchers world wide. The MILC and Fermilab collaborations normally make the configurations available six months following the first publications featuring new high-value ensembles. JNS: *MILC needs to provide the exact statement of their official policy* In the long term, we will make copies of the HISQ ensembles freely available via the NERSC Gauge Connection science portal.